実生活tweet に対する局面の階層的推定法に関する研究

実生活 tweet に対する
局面の階層的推定法に関する研究

筑波大学
図書館情報メディア研究科

2014 年 3 月

山本修平

目次

図目次

表目次

第 1 章

序論

1.1 背景

現在，知識共有コミュニティサイトやブログ，マイクロブログなど，多くの情報共有サービスが存在している．tweet と呼ばれる短文を投稿する Twitter[*1]は，最も広く普及しているマイクロブログの一つであり，2013 年現在，20 億以上ものユーザが 40 億を超える tweet を日々投稿している [7]．ユーザは自らの経験や意見，また日常生活でのイベントなど，身近な「今」を投稿しているため，他のユーザにとっても最新かつ有益な tweet が多い．例えば，電車の遅延情報は交通機関を利用するユーザに役立ち，スーパーの特売情報はこれから買物に出かけようとしているユーザを支援できる．これらのような地域性が高く新鮮かつ，他のユーザに有益な tweet を，本論文では「実生活 tweet」と呼ぶ．

実生活 tweet が実際にユーザの生活を支援した例として，2011 年 3 月に起きた東日本大震災が知られている [25]．地震が起きた直後，被災地周辺では断水や食料供給の不足，交通機関の運行停止など，大きな混乱が生じた．その際，給水や食料配布が行われる場所，電車やバスの運行情報について記述された有益な tweet が数多く投稿され，多くの生活者を支援したと報告されている．

このように，ユーザにとって有益な実生活 tweet は，Twitter に数多く投稿されている．一方で，実生活 tweet 以外の tweet も少なくない．特に，「ありがとう」や「そうなんだ」，「なるほど」といった，誰かの投稿に対する相槌や共感などの，ユーザの生活を直接支援しない tweet が多い．このような tweet の混在は，実生活 tweet の発見を妨げる原因となっている．

[*1] http://twitter.com

1.2　実生活の局面

実生活 tweet は，様々な局面に対応している．例えば，「電車が来ない」という tweet は生活の中の「交通」の局面に対応し，これから電車に乗ろうとしているユーザを支援できる．「今日は全商品半額です」という tweet は「消費」の局面に対応し，買物に行こうとしているユーザを支援できる．本研究では，Wikipedia の「地域コミュニティ」[*2]と「生活」[*3]を参考に，人々の生活を表 1.1 に示す 14 の局面に整理する [26].

1.3　本研究の目的

本研究では，実生活 tweet を抽出するため，未知の tweet に対して表 1.1 に示す局面を付与する，局面の階層的推定法を提案する．階層的推定法は，第一段階で大量の tweet から潜在的ディリクレ配分法（LDA）を用いてトピックを抽出し，第二段階で少量のラベル付き tweet

表 1.1　実生活の局面

局面	典型的な単語
服飾	衣服, 服装, 着る, 装飾, 化粧, 理髪, 衣装 ...
交流	約束, 出会い, 招待, 友人, 誘い, 勧誘, 飲み会 ...
災害	洪水, 竜巻, 地震, 火事, 津波, 二次災害 ...
食事	料理, 外食, 食べ物, レストラン, ジャンクフード ...
行事	祭り, 冠婚葬祭, 日程, 開催日, 学園祭, 文化祭 ...
消費	購入, 買う, 注文, 安売り, 特売, ショッピング ...
健康	風邪, 体調, 怪我, 痛み, 健康法, 病気予防 ...
趣味	余暇, 娯楽, おもちゃ, 音楽, テレビ, ゲーム ...
居住	掃除, 家具, 洗濯, 住まい, 隣人, アパート ...
地域	観光, 地域情報, 地理情報 ...
学校	勉強, 宿題, 課題, 試験, テスト, 資格, 研究 ...
交通	電車, バス, 飛行機, 時刻表, 渋滞, 混雑, 遅延 ...
気象	天気, 気温, 湿度, 風, 花粉, 雨量, 空模様 ...
労働	アルバイト, 研修, 就職活動, 営業, 仕事 ...

[*2] http://ja.wikipedia.org/地域コミュニティ
[*3] http://ja.wikipedia.org/生活

を用いて，トピックと局面の関連度を算出し，局面毎に閾値を超えた関連度を持つトピックと対応関係を構築する．入力された未知の tweet をトピックに展開し，関連度を用いて局面毎にスコアを算出する．スコアに基づいて未知の tweet に対して局面を推定するが，実生活 tweet は複数の局面を付与した方が適切な場合もある．例えば，「激しい雨のため，電車が遅延しています」という tweet は，「激しい雨が降っている」ことと，「電車が遅延している」ことについて言及している．tweet の主題は「電車の遅延」であるが，この tweet はすぐに外の状況を確認できないユーザに対して「雨が降っている」ことを伝えることも可能である．ユーザは，現在の活動状況に適した局面を指定して，必要とする情報を選択して入手するであろうから，この tweet には「交通」だけでなく，「気象」も付与しておくことが望ましい．tweet に複数の局面を付与するため，算出したスコアから閾値を決定し，閾値を超えたスコアを持つ局面を tweet に対して推定する．

1.4　本論文の構成

本論文の構成は次の通りである．第 2 章では，関連研究について概観し，本研究の位置づけを示す．第 3 章では，実生活 tweet を抽出するために，局面の階層的推定法について説明する．第 4 章で，人出判定によって作成した正解データを用いて，提案手法の推定精度について評価し，5 章で考察する．第 6 章で本論文のまとめと，今後の課題について述べる．

第 2 章

関連研究

2.1 経験マイニングに関する研究

実生活 tweet は，ユーザの経験や知識，或いは地域に特有の情報を含んでいる．文書から経験情報を抽出する方法として，経験マイニングに関する研究がいくつか行われている．Kurashima ら [10] は，人間の経験を {状況，行動，主観} からなる情報と捉え，文章中から {時間，空間，動作，対象，感情} を自動抽出する手法を述べている．Inui ら [6] は，人間の経験を {時間，極性，話者態度} の観点から，{トピック，経験主，事態表現，事態タイプ，事実性} の各項目に索引付する枠組みを提案している．池田ら [24] は，ユーザの体験表現を 20 種類の品詞の組合せルールとして定義し，ルールに適合する文章があった場合に，体験表現として抽出している．Hattori ら [5] は，ソーシャルメディアに投稿される耳寄り情報を抽出するために，LDA によって各トピックにクラスタリングしたコメントから，経験マイニングを用いて経験情報を抽出した後に，耳寄りキーワードを含むコメントを抽出している．これらの文章の構造に着目した経験マイニングに関する研究は，ブログなどの長い文章に対して効果的であるが，Twitter に投稿される記事のような，非常に短い文書に対しては有効に機能しないと考えられる．Twitter に投稿される記事は，頻繁に主語や目的語が省略され，経験マイニングをより難しくしている．本研究は Twitter を対象に，ユーザの経験だけでなく知識や地域特有のイベントを抽出することを目的としている．

2.2 Twitter からの情報抽出に関する研究

Twitter からの有益な情報を抽出する研究は，数多く行われている．Sakaki ら [17] は，Twitter ユーザをセンサーとみなし，地震などの現実世界で起きるイベントを発見する手法を明らかにしている．Mathioudakis ら [14] は，収集した tweet からバーストキーワードを見つけ出し，キーワードの共起を用いてクラスタリングを行い，リアルタイムに変動するトレン

ドを発見することを目指している．Zhao ら [23] は，Twitter に投稿された情報要求に関する tweet を抽出し，ユーザの情報要求から現実世界のイベントやトレンドを発見する手法を提案している．Wang ら [18] は，ユーザに対して tweet を推薦する手法を述べている．過去の tweet からユーザの興味を推定し，入力された tweet に対するユーザのスコアを求め，スコアに基づいて Support Vector Regression を用いてユーザをランキングすることにより，tweet と関連深いユーザの抽出をしている．Li ら [11] は，Twitter に投稿される固有表現を教師なし学習によって抽出する手法を提案している．Tweet を単語列に分解し，相互情報量を用いて単語列のスコアを算出した後に，固有表現は他の固有表現と共起し易いという仮説に基づいて，特定期間に投稿された tweet の単語列間の共起度から，固有表現らしい単語列をランキングしている．本研究は未知の tweet に対して実生活の局面を推定することによって，有益な実生活 tweet を抽出することを目的としているため，これらの Twitter に関する研究とは異なる．

2.3　トピックモデルに関する研究

トピックモデルに関する研究では，Blei ら [1] によって提案された，潜在的ディリクレ配分法（LDA）が広く知られている．LDA とは，一つの文書に複数のトピックが存在すると仮定した確率的トピックモデルであり，それぞれのトピックがある確率を持って文書上に共起するという考えのもと，各トピックの確率分布を導出する教師なし学習モデルである．Riedl ら [16] は，LDA を用いてテキストを話題毎に分割する手法を述べている．LDA によって得られた各トピック中の単語の生起確率から，文書中の単語をトピック ID に変換する．文の境界の前後に一定の文章数の窓を設定し，各窓毎にトピックの出現頻度のベクトルを算出することにより，トピックの変換点を検出している．Zhang ら [22] は，LDA を用いてアーティストの推薦をする手法を提案している．ユーザの選考アーティストを特徴量として生成したトピック集合と，アーティストのコミュニティに所属するユーザを特徴量として生成したトピック集合を用いて，アーティスト間の類似度，ユーザ間の類似度を算出し，精度だけでなく意外性のあるアイテムの推薦も目指している．Weng ら [20] は，Twitter 上で影響力のあるユーザの発見のために，LDA を用いてユーザのトピック同定をしている．トピック毎にユーザのネットワークを構築し，各ネットワークに対して PageRank を拡張した TwitterRank を適用することで，ユーザをランキングする手法を提案している．本研究は，生成したトピックと局面の対応関係を構築し，トピックと局面の関連度とトピック中の単語の生起確率を用いて，未知の tweet に対して複数ラベルを推定することに特徴がある．

2.4　本研究の位置づけ

これまでに，Twitter を対象に有益な情報を抽出する研究は数多くなされているが，本研究では，人々の生活に有益な実生活 tweet を抽出することを目的としていることに特徴がある．また，実生活 tweet は人々の経験だけでなく，知識や地域特有のイベントも含んでいるため，文書から経験情報を抽出する経験マイニングに関する研究とは異なる．トピックモデルは，バーストキーワードやユーザのトピック推定，アイテム推薦などに用いられる研究が多いが，本研究では，トピックモデルによって生成したトピックと局面の対応関係を用いる階層的推定法を提案する．階層的推定法はトピックと局面の関連度を用いて，未知の tweet に対して複数の局面を推定する．

第 3 章

局面の階層的推定法

本章ではまず，階層的推定法の概要とそのアプローチについて説明する．次に，階層的推定法の第一段階であるトピックの抽出方法と，第二段階のトピックと局面の対応関係の構築方法について詳述する．次に，トピックと局面の対応関係におけるトピック数の最適化方法について解説し，最後に未知の tweet に対する局面の推定方法について述べる．

3.1 局面推定のアプローチ

実生活 tweet は表 1.1 に提示したように，様々な局面を含んでおり，局面に関連する全てのキーワードを列挙することは困難である．また，経験マイニングで用いられているルールベースの解析手法は，Twitter に投稿される記事が短く省略が多いことから，実生活 tweet の抽出を行うことは難しいと考えられる．

本研究で提案する階層的推定法の概要について，図 3.1 に示す．第一段階では，LDA を用いて大量の tweet からトピックを抽出する．LDA は大量の文書集合をクラスタリングするための，教師無し学習モデルであるため，教師（正解）ラベルを必要としない特徴がある [1]．第二段階では，局面ラベルが付与された少量の tweet を用いて，トピックと局面の対応関係を構築する．入力された tweet から抽出した単語の各トピックにおける生起確率と，トピックと局面の対応関係を用いて，局面毎にスコアを算出し，スコアが閾値を超えた局面を未知の tweet に対して推定する．

局面とトピックの対応関係を構築することによって，各局面が対象とする話題をトピックによって構成することが可能となる．また，LDA で抽出するトピックは，似たような話題で使用される単語を同一のトピック中に高い生起確率でクラスタリングできる．このため，未知の tweet には出現するが正解データに出現しなかった単語についても，トピック中の単語の生起確率をスコアを算出する際の特徴量として用いることができる．以上のことから，階層的推定法は少量の正解データでトピックと局面の対応関係を学習できるため，幅広い話題を対象とす

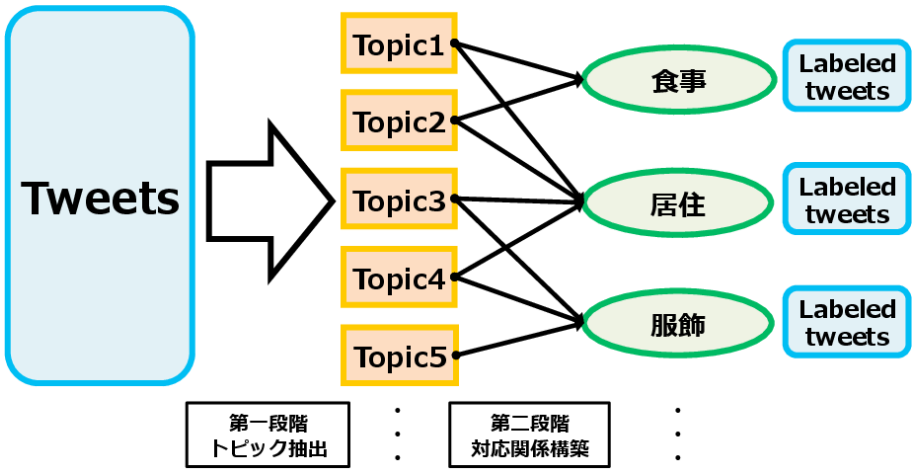

図 3.1　階層的推定法の概要

る実生活 tweet を網羅的に抽出することが期待できる．

3.2　LDA を用いたトピックの抽出方法

大量の tweet を LDA の入力とし，トピックを抽出する．LDA とは，一つの文書に複数のトピックが存在すると仮定した確率的トピックモデルであり，それぞれのトピックがある確率を持って文書上に共起するという考えのもと，各トピックの確率分布を導出する教師なし学習モデルである．

LDA のグラフィカルモデルを図 3.2 に示す．各文書はトピック分布 θ を持ち，文書上の各単語について θ に従ってトピック z が選ばれ，そのトピック z に対応する単語分布 ϕ に従って，単語 w が生成される．K はトピック数，D は文書数，N_d は文書 d 上の単語の出現回数を表しており，トピック分布 θ は各文書毎に，単語分布 ϕ は各トピック毎に，単語 w とその単語のトピックを表す z は，各単語の出現する位置ごとに生成される．α はパラメータ θ が従うディリクレ分布のハイパーパラメータ，β はパラメータ ϕ が従うディリクレ分布のハイパーパラメータを示す．実際に観測される変数は単語 w である．

LDA における文書の生成過程は，以下のような手順である．

1. 各トピック k について，ディリクレ分布に従って単語分布 ϕ_k を生成
   $\phi_k \sim Dir(\beta)$
2. 各文書 d について，ディリクレ分布に従ってトピック分布 θ_d を生成
   $\theta_d \sim Dir(\alpha)$
3. 文書 d における各単語の位置 n について，
   (a) 文書 d における各単語の位置 n について，多項分布に従ってトピック $z_{d,n}$ を生成

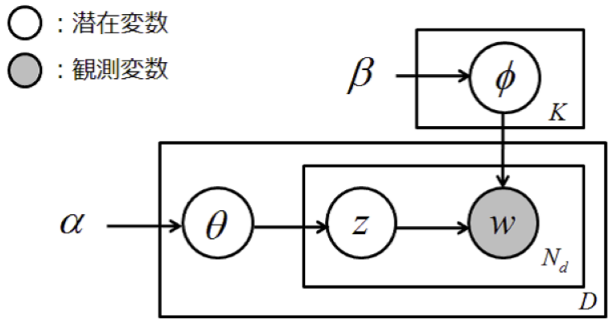

図 3.2　LDA のグラフィカルモデル

$z_{d,n}$〜$Multi(\theta_d)$

（b）文書 d における各単語の位置 n ついて，多項分布に従って単語 $w_{d,n}$ を生成
$w_{d,n}$〜$Multi(\phi_{z_{d,n}})$

ここで，$Dir(\cdot)$ はディリクレ分布，$Multi(\cdot)$ は多項分布を表す．

LDA は，教師なし学習によって文書ー単語行列からトピック集合 Z を推定する必要がある．本研究では，ギブスサンプリングを用いて推定する [4]．ギブスサンプリングの更新式は，以下の式で与えられる．

$$P(z_i|Z_{-i},W) = \frac{n^d_{-i,j}+\alpha}{n^d_{-i,\cdot}+T\alpha} \cdot \frac{n^v_{-i,j}+\beta}{n^{\cdot}_{-i,j}+W\beta} \tag{3.1}$$

ここで，Z_{-i} はトピック集合全体から i 番目のトピックを除いたものを表している．$n^d_{-i,j}$，$n^d_{-i,\cdot}$，$n^v_{-i,j}$，$n^{\cdot}_{-i,j}$ は，それぞれ位置 i の情報を除外した場合の，文書 d においてトピック j が割り当てられた頻度，文書 d において単語が生成された頻度，トピック j から単語 v が生成された頻度，文書全体においてトピック j が割り当てられた頻度を表す．また，T はトピック数，W は語彙数を表している．

ギブスサンプリングによって得られたサンプルから，各文書におけるトピック分布 θ と，各トピックにおける単語分布 ϕ の予測分布を算出する．文書 d においてトピック k が生成される確率の推定量 $\hat{\theta^k_d}$，トピック k が選択されたときに単語 w が生成される確率の推定量 $\hat{\phi^w_k}$ は，以下の式から算出される．

$$\hat{\theta^k_d} = \frac{N_{d,k}+\alpha}{N_d+T\alpha} \quad , \quad \hat{\phi^w_k} = \frac{N_{k,w}+\beta}{N_k+W\beta} \tag{3.2}$$

本研究では，第一段階で抽出するトピックとして，トピック k における単語 w の生起確率である $\hat{\phi^w_k}$ を抽出する．トピック k における単語 w の生起確率を用いて，トピックと局面の対応関係を構築する．

3.3　トピックと局面の対応関係の構築方法

大量の tweet を LDA の入力とし，予めトピックを生成しておく．抽出するトピックは数百程度である．トピックと局面の対応関係を構築するため，局面がラベル付けされた少量の tweet を用意する．用意された tweet に対して形態素解析し[*1]，得られた語彙集合を W とする．形態素解析には，日本語形態素解析器 MeCab[9] を用いる．ここで，抽出した全ての単語に対して，局面を特徴付ける単語の重みとして情報利得 [21] を用いる．単語 w の情報利得 $IG(w)$ は，以下の式で与えられる．

$$IG(w) = H(A) - (P(w)H(A|w) + P(\bar{w})H(A|\bar{w})) \tag{3.3}$$

ここで，A は全ての局面を意味する．$P(w)$ は全ての tweet の中で単語 w が出現する確率，$P(\bar{w})$ は全ての tweet の中で単語 w が出現しない確率である．$H(A|w)$ は単語 w が出現するときの，全局面 A における条件付きエントロピー，$H(A|\bar{w})$ は単語 w が出現しないときの，全局面 A における条件付きエントロピーである．$IG(w)$ の値が高いとき，単語 w は良い特徴であることを意味する．

各局面 a における単語 w の重要度として，単語の生起確率 $p(a, w)$ を算出する．本論文では，マルチラベルが付与されているデータから，単語の生起確率を算出できる Labeled LDA[15] を用いる．Labeled LDA（L-LDA）は，Ramage らによって LDA を教師あり学習へ拡張したトピックモデルである．L-LDA は文書に予め付与されているラベルを，その文書の内容を表すものと捉えることで，潜在トピックの抽出における教師ラベルとして利用することを考えたモデルである．L-LDA は文書に複数のラベルを付与した状態で，クラス別に単語の生起確率を算出できる．

トピック t と局面 a の関連度 $R(a, t)$ は，

$$R(a, t) = \sum_{w \in W} IG(w) * p(a, w) * p(w, t) \tag{3.4}$$

で算出する．ここで，$p(w, t)$ は，LDA を用いて抽出したトピック t における単語 w の生起確率である．この式は，局面 a の単語の生起確率とトピック t の単語の生起確率を用いて，局面 a とトピック t の関連度を算出する．

関連度を 0 から 1 の範囲とするために，正規化する．ここでは，以下の式（3）に示す，各局面で正規化した関連度 $\hat{Ra}(a, t)$ と，各トピックで正規化した関連度 $\hat{Rt}(a, t)$ を用意する．

[*1] 実際には，形態素解析の結果に基づいて，名詞と動詞，形容詞のみを使用する．

$$\hat{Ra}(a,t) = \frac{R(a,t)}{\sum_{t \in T} R(a,t)} \quad , \quad \hat{Rt}(a,t) = \frac{R(a,t)}{\sum_{a \in A} R(a,t)} \tag{3.5}$$

ここで，T は LDA で抽出した全てのトピック，A は全ての局面である．$\hat{Ra}(a,t)$ は，局面がどのトピックによって支持されているかを表す指標であり，$\hat{Rt}(a,t)$ は，トピックがどの局面を支持しているかを表す指標である．

関連度 $\hat{Ra}(a,t)$ を用いてトピックと局面の対応関係を構築する．局面によって，対応関係にも様々な傾向があることが考えられる．局面とトピックの対応関係の例を図 3.3 に示す．この図は，交通の局面はトピック 2 から 0.79 の高い関連度を持っており，消費の局面はトピック 1 とトピック 3 から 0.39 と 0.53 のやや高い関連度を持っている．このような場合，交通の局面はトピック 2 と対応関係を構築し，消費の局面はトピック 1 と 3 の二つのトピックと対応関係を構築する．以上のような対応関係を全てのトピックと全ての局面の間で構築するために，局面毎に閾値を決定し，関連度 $\hat{Ra}(a,t)$ が閾値を超えたトピックと対応関係を構築する．局面 a と対応関係を構築するトピック集合 T_a は，媒介変数 d を用いて，

$$T_a = \{t | \hat{Ra}(a,t) > \max_{t \in T}(\hat{Ra}(a,t)) - \sigma(\hat{Ra}(a,T)) * d\} \tag{3.6}$$

とする．ここで，$\sigma(\hat{Ra}(a,T))$ は，局面 a の全トピック T に対する関連度の標準偏差である．d はトピックとの対応関係を調整するためのパラメータであり，d を大きくすることで，より多くのトピックが局面に関連付けられる．トピックが局面に関連付く度合いは局面毎に異なることから，第 4 章では d を変化させて，適合率，再現率，F 値を評価し最適な値を求める．

3.4　局面の推定方法

未知の tweet の局面を推定するため，第 3.3 節で構築したトピックと局面の対応関係を用いる．局面の推定方法の概要を図 3.4 に示す．tweet から単語を抽出し，トピック中の各単語の生起確率と，トピックと局面の関連度を用いて，局面毎にスコアを算出する．未知の tweet tw と各局面 a のスコア $S(tw,a)$ は，以下の式で算出する．

$$S(tw,a) = \sum_{t \in T_a} \sum_{w \in W_{tw}} p(w,t) * \hat{Ra}(a,t) * \sigma(\hat{Rt}(A,t)) \tag{3.7}$$

ここで，W_{tw} は未知の tweet tw から抽出した単語集合，$p(w,t)$ はトピック t における単語 w の生起確率である．$\sigma(\hat{Rt}(A,t))$ は，トピック t の全局面 A に対する関連度 $\hat{Rt}(A,t)$ の標準偏差である．$\sigma(\hat{Rt}(A,t))$ の値が高いとき，トピック t は特定の局面に対して強く支持している．$\sigma(\hat{Rt}(A,t))$ の値は，あるトピック t が全ての局面に対して支持している度合いを表す指標となっている．

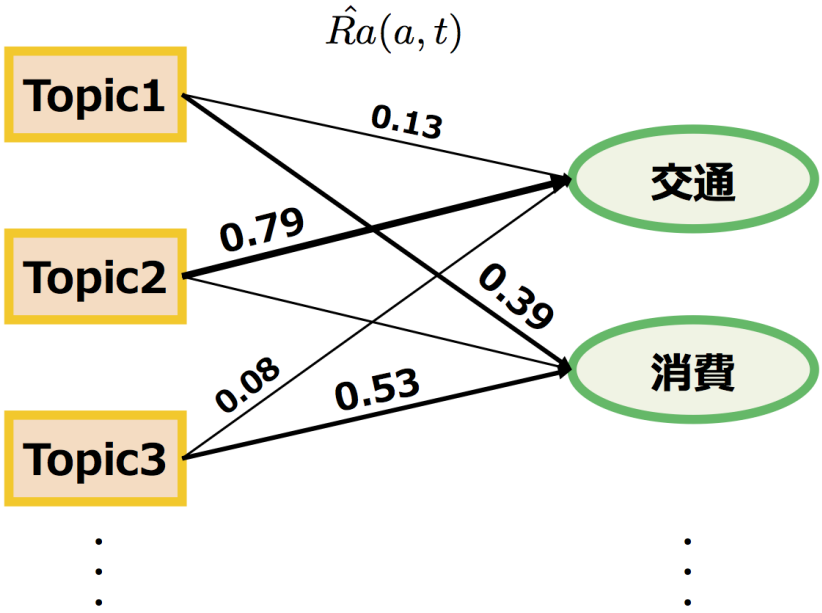

図 3.3　対応関係の例

算出されたスコアが高いほど，未知の tweet に対して推定すべき局面である．本論文で目的としているマルチラベルを実現するためには，上位 k 局面を付与するのだが，k を動的に定める必要がある．

本論文では，算出されたスコアの平均値と標準偏差を用いて各スコアを正規化し，閾値 r を超えたスコアを持つ局面を，未知の tweet に対して推定する．未知の tweet tw に対して推定する局面集合 A_{tw} は，

$$A_{tw} = \left\{ a \middle| \frac{S(tw,a) - E(S(tw,A))}{\sigma(S(tw,A))} > r \right\} \tag{3.8}$$

とする．ここで，$E(S(tw,A))$ と $\sigma(S(tw,A))$ は，スコア $S(tw,a)$ の全局面 A における平均値と標準偏差である．各スコアについて，平均値との差を標準偏差で除すことによって，0 を基準とした値に正規化される．スコアが閾値 r よりも大きければ，そのスコアを持つ局面は未知の tweet に対する局面として推定する．

3.5　トピック数の最適化方法

図 3.1 より，階層的推定法は第一段階で LDA を用いてトピックを抽出し，第二段階でトピックと局面の対応関係を構築する．第 3.3 節で説明したように，関連度に基づいてトピックと局面の対応関係が構築されることから，LDA で生成するトピック数によって，局面と結びつくトピックが大きく異なる．

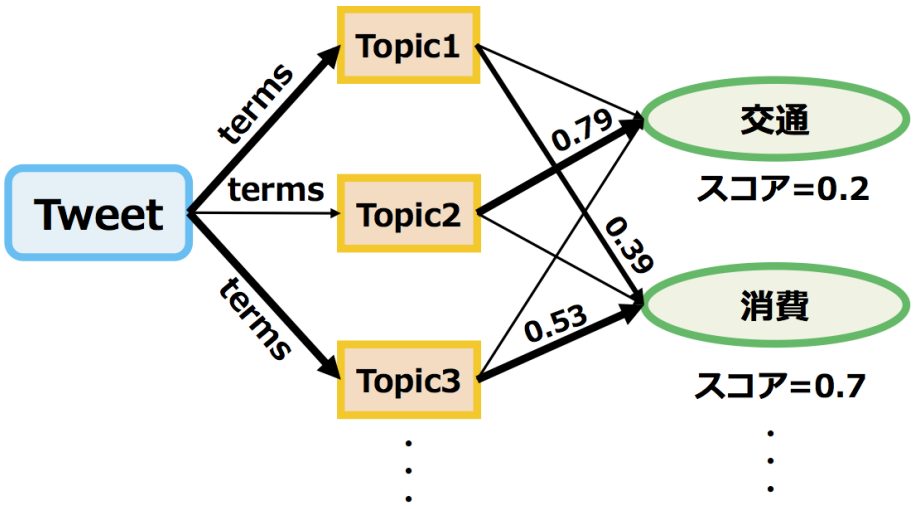

図 3.4　局面の推定方法

最適なトピック数を決定するため，Jensen-Shannon Divergence（JS Divergence）[13] を用いて，ある一つの局面と他の局面との類似度を計算する．二つの局面の確率分布が同じである場合，JS Divergence は 0 となる．本論文の場合は，局面間の確率分布 $\hat{Ra}(a,t)$ が異なっている方が望ましい．そのため，各局面間の JS Divergence の合計値を最大とするトピック数を最適であるとした．JS Divergence の合計値 JS_{sum} は，以下の式で求められる．

$$JS_{sum} = \sum_{(\forall p, \forall q) \in A} JS(\hat{Ra}(p,*)||\hat{Ra}(q,*)) \tag{3.9}$$

$$JS(P||Q) = \frac{1}{2}\left(\sum_{t \in T} P(t) \log \frac{P(t)}{R(t)} + \sum_{t \in T} Q(t) \log \frac{Q(t)}{R(t)}\right)$$

ここで，$R(t)$ は確率分布 $P(t)$ と $Q(t)$ の平均であり，$R(t) = \frac{P(t)+Q(t)}{2}$ で与えられる．

第 4 章

実験と評価

評価実験では，提案手法の適合率，再現率，F 値を評価する．提案手法が有効に機能することを分析するため．Naive Bayes をマルチラベリングへ拡張した手法と比較する．また，構築したトピックと局面の対応関係について詳細に分析することで，提案手法が有効に機能する局面と機能しない局面の違いを明らかにする．

4.1 データセットとパラメータ設定

4.1.1 データセット：トピック抽出に用いる tweet

階層的推定法の第一階層における LDA を用いたトピック抽出のため，2012 年 4 月 15 日から 2012 年 8 月 14 日の間に，日本語で Twitter に投稿された tweet を Search API[*1]を用いて収集した．その中から，tweet のロケーション情報に「京都」或いは「Kyoto」と入力されている tweet を使用した．以上の条件により評価に使用する tweet は，2,390,553 件であった．

4.1.2 データセット：実生活 tweet

トピックと局面の対応関係を構築するため，人手によって局面がラベル付けされた tweet を用意した．1,500 件の tweet に対して，第一著者（実験者 A）と他 2 名（実験者 B 及び C）の合計 3 名で，各 tweet に対して局面を付与する人手判定を行った．実験者にはガイドラインとして，表 1.1 に示す各局面に含まれる典型的な単語と，その局面に分類される tweet の例（各局面 1 件ずつ）と，それが分類された理由を提示した．人手判定では，各 tweet に対して適切な局面として第一，第二，第三候補まで付与することとした．いずれの局面にも適さないと判断した場合は，「非実生活」を付与することとした．なお，用意した 1,500 件の tweet は，い

[*1] https://dev.twitter.com/docs/api/1/get/search

ずれもロケーション情報に「京都」或いは「Kyoto」と表記されたものである．また，3 名の実験者はいずれも「つくば市」在住の大学生である．

人手判定の結果, 第一候補に分類された局面について, 実験者間の一致度を κ 値 [2] によって評価した．実験者 A と実験者 B の κ 値は 0.687, 実験者 A と実験者 C の κ 値は 0.595, 実験者 B と実験者 C の κ 値は 0.576 となった． κ 値の平均は 0.619 であり，高い一致（substantial）であった．

各 tweet に対して適切な複数の局面をラベル付けするため，3 名の人手判定の結果を用いる．tweet tw に対して正解となる局面集合 AC_{tw} は，

$$AC_{tw} = \{a | Uscore(tw, a) \geq 10\} \tag{4.1}$$

とする．ここで，$Uscore(tw, a)$ は，実験者が tweet tw に対して，局面 a を第何候補に選択したかを合計した値であり，以下の式で求められる．

$$Uscore(tw, a) = \sum_{u \in U} candidate(tw, a, u) \tag{4.2}$$

ここで, U は全ての実験者を表し，$cnadidate(tw, a, u)$ は，tweet tw に対して実験者 u が局面 a を分類した候補番号である．$Uscore(tw, a)$ の最小値は, 実験者 3 名が同じ局面を第一候補に選択した場合であり，$candidate(tw, a, u) = 1$ となるため，$Uscore(tw, a) = 3$ となる．最大値は，実験者が特定の局面をいずれの候補にも選択しなかった場合に $candidate(tw, a, u) = 4$ とし，$Uscore(tw, a) = 12$ となる．

以上の処理によって，人手判定した 1,500 件の tweet に対して，各 tweet に付与された局面の数を集計した結果を表 4.1 に示す．最も多いラベル数は 3 で，820 件の tweet が存在する．ラベル数が 6 ある tweet は，11 件存在する．1,500 件の tweet にラベル付けされた局面の数を集計した結果を，表 4.2 に示す．服飾の局面は，1,500 件の tweet の中で合計で 181 件ラベル付けされている．1,500 件の tweet に対する全てのラベル数は 5,092 件となっており，一つの tweet に対して平均 3.39 件のラベルが付与されている．

評価実験では，いずれの局面にも属さない「非実生活」についても一つのクラスとしてトピックと対応関係を構築し，非実生活を推定できるか否か評価する．

表 4.1　ラベル数別の tweet 数

ラベル数	1	2	3	4	5	6	合計
tweet 数	1	111	820	442	115	11	1,500

表 4.2　人手判定の結果，正解ラベルとして付与された局面の数

局面	ラベル数
服飾	181
交流	379
災害	86
食事	287
行事	311
消費	435
健康	177
趣味	348
居住	213
地域	432
学校	195
交通	169
気象	226
労働	262
非実	1,392
合計	5,092

4.1.3　パラメータ設定

LDA は，事前にいくつかのパラメータを設定する必要がある．関連研究 [4] を参考に，ハイパーパラメータである α は $\frac{50}{|T|}$，β は 0.1 とした．$|T|$ は LDA で生成するトピック数である．イテレーション回数は，予備実験の結果から安定した値が得られる 100 とした．

LDA で生成するトピック数は，第 3.5 節で説明した JS_{sum} の値が最大値となるトピック数を選択する．トピック数を 50，100，200，500，1,000 と変化させ，各トピック数で JS_{sum} を算出した結果を，表 4.3 に示す．JS_{sum} が最大となったトピック数は 200 であったことから，最適なトピック数は 200 とし，トピックと局面の対応関係を構築した．

表 4.3　各トピック数における JS_{sum}

トピック数	50	100	**200**	500	1,000
JS_{sum}	114.54	129.64	**135.94**	134.97	129.32

4.2　実験方法

4.2.1　評価尺度

提案手法により，tweet に対して正確に局面が推定できているかを評価する．提案手法の有効性を議論するには，推定した局面がどれだけ正解しているかという正確性と，全ての正解のうちどれだけ提案手法で局面を推定できたかという網羅性の 2 つの観点からの評価が必要となる．本研究では，正確性を適合率（$Precision$），網羅性を再現率（$Recall$），適合率と再現率の調和平均である F 値（$F-measure$）によって提案手法の推定精度を評価する [12]．それぞれの計算方法について，以下に示す．

$$Precision = \frac{\text{推定した正解局面数}}{\text{推定した局面数}} \tag{4.3}$$

$$Recall = \frac{\text{推定した正解局面数}}{\text{全ての正解局面数}} \tag{4.4}$$

$$F-measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.5}$$

実験では，分割交差検定 [8] によってトピックと局面の対応関係を構築，及び推定精度の評価を行う．正解データの分割数を 10 とし，9 割のデータでトピックと局面の対応関係構築し，残りの 1 割を用いて推定精度を評価する．以上の操作を 10 回繰り返し，各評価値の平均値を算出する．

4.2.2　比較手法

提案手法の有効性について検証するため，本論文では，文書に対してマルチラベル分類ができる Naive Bayes Multi-label Classification（NBML）[19] を比較対象とする．NBML は Wei らによって，Naive Bayes[3] をマルチラベル分類へ拡張された手法である．Naive Bayes は，テキスト中に含まれる単語が互いに独立に発生したものであるという仮定をおき，それらの単語が出現したときの文書のクラスへの所属確率をベイズの定理により求め，所属確率が最も高いクラスへ文書を分類する手法である．NBML は，Naive Bayes によって算出したクラス別の所属確率から平均値を求め，平均値を超える所属確率を持つクラスへ文書を分類することで，マルチラベリングを実現している．

Naive Bayes は単語の生起確率を算出する際に，各文書に対して一つのラベルが付与されていることを前提としているため，第 4.1.2 節で説明したような複数ラベルが付与された文書に

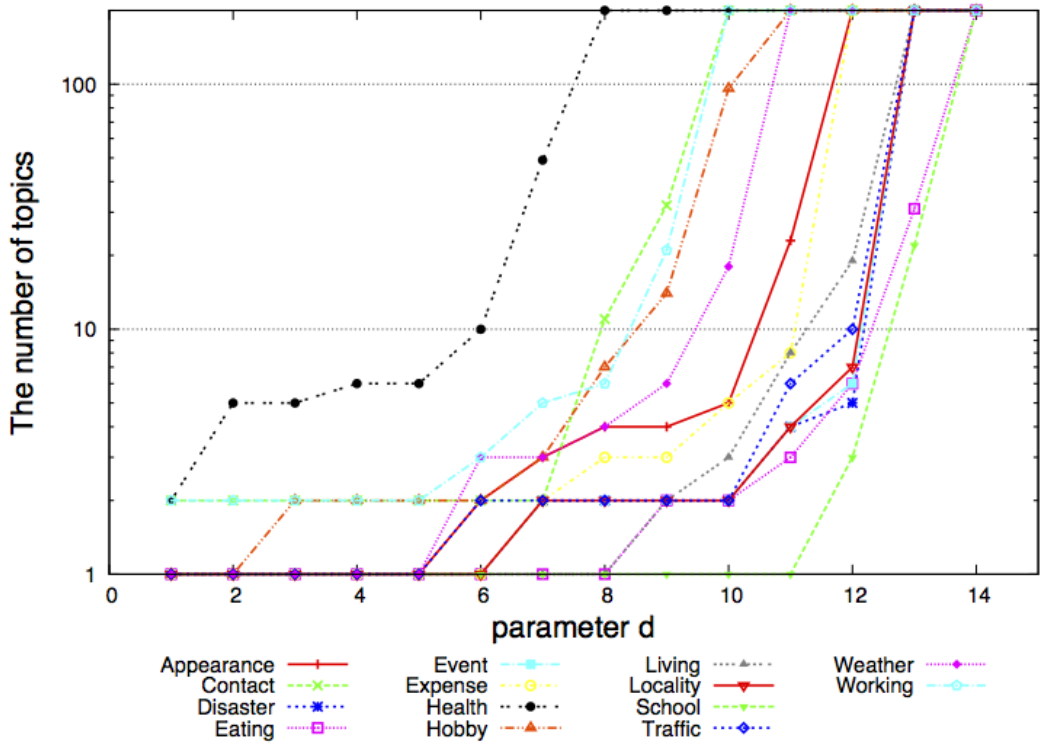

図 4.1　各局面に結びつくトピック数の遷移

対応していない．そこで，本論文では文書に付与されているラベルを一つにするため，ラベル毎に同じ文書を複製することで対応する．つまり，ラベルの数だけ文書を複製するため，表 4.2 に示しているように，全てのラベルを合計した 5,092 件の tweet によって学習する．

4.3　実験結果

4.3.1　局面に対応するトピックの推定結果

全ての局面に結びつくトピック数が 200 となるまで，パラメータ d を 1 から増加していく．各局面に対して結びつくトピック数の遷移を図 4.1 に示す．全ての局面で，パラメータ d の増加にともなって，結びつくトピック数が増加している．学校の局面は $d \leq 11$ の範囲で結びついているトピック数が一つである．$d = 14$ となったとき，全ての局面で結びついているトピック数が 200 となり，トピックと局面の対応関係は完全二部グラフとなる．

4.3.2　推定の精度

提案手法を用いて適合率，再現率，F 値を評価した．提案手法について式 3.9 に示した $r = 0$ としたときの，学校と交通の局面のパラメータ d の増加にともなう各評価値とトピック数の変

化を，それぞれ図 4.2 と図 4.3 に示す．学校の局面では，d の増加にともなって適合率は増加し，$d \geq 12$ から急激に減少に転じている．再現率は $d = 14$，F 値は $d = 10$ で最大となっている．交通の局面では，いずれの評価値についても $d = 6$ で増加している．再現率は $d \geq 7$ から減少しているものの，$d = 11$ で再び増加している．また，F 値も $d = 11$ で最大となっている．

F 値が最大値となったときの適合率，再現率，F 値を表 4.4 に示す．提案手法で $r = 0$ の場合は TPE0，$r = 1$ の場合は TPE1，比較手法は NBML と表記した．太字で表記している数値は，各手法を比較したときの最大値を表している．服飾の局面では，適合率の最大値は 0.83 で NBML，再現率の最大値は 0.57 で TPE0，F 値の最大値は 0.53 で TPE0 となっている．全ての局面についてマクロ平均をとった結果，各手法で最大値を示したのは，適合率と F 値では NBML，再現率で TPE0 であったが，F 値における差異は僅かであった．

提案手法と比較手法について，各 tweet に対して推定したラベル数を集計した結果を表 4.5 に示す．TPE0 では最頻値となったラベル数は 5，TPE1 では 0，NBML では 2 となった．また，いずれの手法についても，最頻値となるラベル数に向けて徐々に増加し，最頻値を超えてからは減少に転じている．一つの tweet に対する平均ラベル付与数は，TPE0 は 4.28，TPE1 は 2.08，NBML は 2.75 となった．

4.3.3　トピックと局面の対応関係の詳細

トピックと局面の対応関係の詳細を，表 4.6 に示す．表では，局面毎に $\hat{Ra}(a,t)$ の高い上位 5 トピックを抽出し，そのトピック番号と関連度を示している．また各トピックについて，$\sigma(\hat{Rt}(A,t))$ の値によって降順に並び替えたときの順位も示している．服飾の局面に対して最も強く結びつくトピックは，Topic147 であり，その関連度 $\hat{Ra}(a,t)$ は 0.261 である．Topic147 については，順位が 1 位となっており，200 トピックの中で最も $\sigma(\hat{Rt}(A,t))$ の値が高いことを表している．災害，行事，地域，交通，気象，非実生活の局面に対しては，共に Topic62 が最も強く結びついている．同様に，Topic174 や Topic21，Topic152 も多くのトピックに対して強く結びついている．

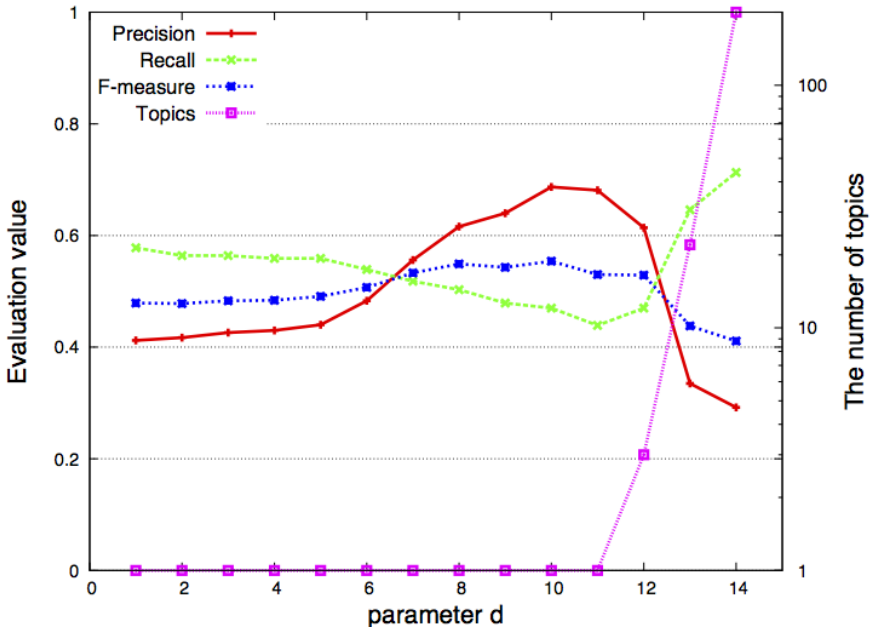

図 4.2　学校の適合率，再現率，F 値

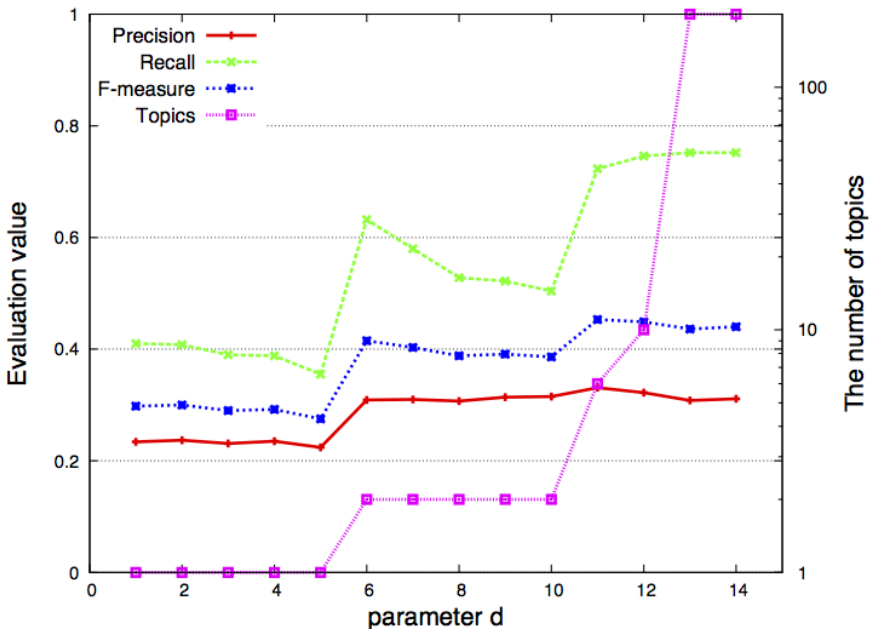

図 4.3　交通の適合率，再現率，F 値

表 4.4　各手法の適合率，再現率，F 値

	適合率			再現率			F 値		
局面	TPE0	TPE1	NBML	TPE0	TPE1	NBML	TPE0	TPE1	NBML
服飾	0.49	0.57	**0.83**	**0.57**	0.46	0.37	**0.53**	0.51	0.51
交流	0.38	0.50	**0.53**	**0.63**	0.38	0.54	0.47	0.43	**0.53**
災害	0.16	0.17	**0.76**	**0.51**	0.47	0.21	0.25	0.25	**0.33**
食事	**0.88**	**0.88**	0.73	**0.60**	0.51	0.51	**0.71**	0.64	0.60
行事	0.35	0.38	**0.56**	0.42	0.34	**0.45**	0.38	0.36	**0.49**
消費	0.43	**0.61**	0.52	**0.49**	0.21	0.46	0.46	0.32	**0.49**
健康	0.28	0.27	**0.76**	**0.60**	**0.60**	0.38	0.38	0.37	**0.50**
趣味	0.39	0.50	**0.57**	**0.67**	0.40	0.44	**0.50**	0.44	0.49
居住	0.46	0.48	**0.71**	**0.47**	0.46	0.41	0.46	0.47	**0.51**
地域	0.81	**0.83**	0.62	0.56	0.50	**0.65**	**0.66**	0.62	0.63
学校	0.68	0.69	**0.81**	0.46	0.42	**0.52**	0.55	0.51	**0.63**
交通	0.33	0.36	**0.82**	**0.72**	0.65	0.50	0.45	0.46	**0.62**
気象	0.33	0.41	**0.81**	0.57	0.40	**0.58**	0.42	0.40	**0.67**
労働	0.36	0.36	**0.56**	**0.58**	0.46	0.35	**0.44**	0.41	0.43
非実	0.94	**0.95**	0.93	0.73	0.21	**0.93**	0.82	0.34	**0.93**
平均	0.48	0.53	**0.70**	**0.57**	0.43	0.49	0.50	0.43	**0.56**

表 4.5　各手法のラベル数別の tweet 数

ラベル数	TPE0	TPE1	NBML	人手判定
1	70	609	165	1
2	122	421	531	111
3	262	215	442	820
4	291	247	243	442
5	478	7	90	115
6	197	0	23	11
7	62	0	6	0
8	17	0	0	0
9	1	0	0	0
平均ラベル数	4.28	2.08	2.75	3.39

表 4.6　各局面に対して高い関連度 $\hat{Ra}$ で結びつくトピックと $\sigma(\hat{Rt}(A,t))$ の順位

	$\hat{Ra}$1 位		$\hat{Ra}$2 位		$\hat{Ra}$3 位		$\hat{Ra}$4 位		$\hat{Ra}$5 位	
	topic ($\hat{Ra}$)	順位	topic ($\hat{Ra}$)	順位	topic ($\hat{Ra}$)	順位	topic ($\hat{Ra}$)	順位	topic ($\hat{Ra}$)	順位
服飾	#147 (0.261)	*1	#89 (0.132)	46	#138 (0.121)	76	#47 (0.089)	34	#60 (0.038)	93
交流	#154 (0.157)	35	#13 (0.157)	31	#191 (0.038)	108	#160 (0.035)	87	#146 (0.031)	154
災害	**#62** (0.504)	99	**#174** (0.236)	98	**#21** (0.078)	106	**#152** (0.076)	100	#69 (0.035)	*5
食事	#132 (0.359)	13	#154 (0.113)	35	#112 (0.067)	*9	#2 (0.043)	40	#89 (0.038)	46
行事	**#62** (0.413)	99	**#174** (0.194)	98	**#21** (0.065)	106	**#152** (0.063)	100	#12 (0.048)	17
消費	#89 (0.234)	46	#132 (0.108)	13	**#62** (0.079)	99	#13 (0.055)	31	#174 (0.038)	98
健康	#34 (0.146)	16	#146 (0.145)	154	#154 (0.116)	35	#158 (0.113)	172	#100 (0.112)	*2
趣味	#4 (0.161)	21	#73 (0.119)	22	#138 (0.050)	76	#188 (0.046)	20	#89 (0.046)	46
居住	#16 (0.248)	*7	#47 (0.073)	34	#176 (0.063)	109	#138 (0.046)	76	#6 (0.037)	119
地域	**#62** (0.408)	99	**#174** (0.199)	98	**#152** (0.064)	100	**#21** (0.064)	106	#4 (0.043)	21
学校	#67 (0.284)	*4	#80 (0.042)	11	#5 (0.035)	19	#176 (0.031)	109	#188 (0.031)	20
交通	**#62** (0.385)	99	**#174** (0.203)	98	**#21** (0.061)	106	#162 (0.060)	12	**#152** (0.058)	100
気象	**#62** (0.192)	99	#47 (0.093)	34	**#174** (0.090)	98	#138 (0.055)	76	#191 (0.038)	108
労働	#39 (0.205)	113	#57 (0.202)	54	#133 (0.089)	41	#193 (0.056)	143	#9 (0.056)	130
非実	**#62** (0.073)	99	#138 (0.042)	76	#132 (0.039)	13	#146 (0.037)	154	**#174** (0.035)	98

第 5 章

考察

5.1 提案手法と比較手法に関する考察

表 4.4 より，再現率では，多くの局面で TPE0（提案手法で $r = 0$）が NBML に比べ高い値を示している．しかし，適合率と F 値では，NBML が提案手法に比べ高い値を示す局面が多いことが分かる．表 4.5 より，一つの tweet に対する平均ラベル数は，TPE0 が 4.28，TPE1 が 2.08，NBML が 2.75 となっている．表 4.1 より，正解データに付与されているラベル数の平均は 3.39 となっている．以上のことから，提案法である TPE0 は正解データに対して過剰な局面を推定する割合も多いが，NBML よりも網羅的に局面を推定できるため，実生活 tweet の抽出に適した手法であると考えられる．

提案手法が正確に正解ラベルを推定した例を，表 5.1 に示す．tweet に正解ラベルとして付与されている局面は気象と健康であり，NBML では気象の局面のみ推定しているが，TPE0 はこの二つの局面を正確に推定することに成功している．tweet 文に着目すると，翌日の天気予報について文章の前半で言及し，文章の後半では天候によって体調を悪くしないための対策について言及している．NBML では，「晴天」や「黄砂」，「気温」の単語が気象の局面に対して高い生起確率であったために，気象に対する所属確率が他の局面に比べて高くなり，一つの局面のみを推定したと考えられる．一方，提案手法では「熱中症」や「水分」といった単語を高い生起確率として持つトピックが，健康の局面に対して高い関連度を持って結びついたために，気象に加えて健康の局面も推定することができたと考えられる．

提案手法が人手による正解ラベルに加えて，別の局面も推定した例を表 5.2 に示す．この例では，tweet に正解ラベルとして付与されている局面は学校と行事であり，NBML ではこの二つの局面を推定することに成功している．これに対して，TPE0 ではこの二つの局面に加えて地域の局面を推定する結果となった．tweet 文に着目すると，「桂川」という単語が存在することが分かる．「桂川」は京都市内を流れる川の名前であるため，この tweet は，「桂川」の近くで勉強会を開催するという内容であることが分かる．したがって，この tweet は勉強会を開

催する内容であるが，その勉強会が特定の地域に限定したイベントであると考えることもできる．このように，人手では思い至らなかった局面についても TPE0 はラベル付けすることができており，網羅性の高いラベリングができていると言える．

5.2　トピック局面の対応関係に関する考察

表 4.4 より，提案手法で 0.5 以上の F 値を示した局面は，服飾，食事，趣味，地域，学校である．表 4.6 より，各局面に対して強く結びつくトピックに注目すると，服飾と学校の局面で $\hat{Ra}1$ 位で結びついている Topic147 と Topic67 は，共に $\sigma(\hat{Rt}(A,t))$ の順位が 1 位と 4 位であり，他のトピックに比べて非常に高いことが分かる．$\sigma(\hat{Rt}(A,t))$ の値の大きさは，特定の局面に対してどれほど強く支持しているかを表す．Topic147 と Topic67 は，それぞれ服飾と学校に対してのみ，高い関連度で結びついていたため，$\sigma(\hat{Rt}(A,t))$ の値が大きくなり，順位も高くなったと考えられる．

食事と趣味の局面に着目すると，$\hat{Ra}1$ 位で結びついているトピックの順位は，Topic147 と Topic67 に比べると低いが，全てのトピックが 200 ある中では高い順位であることが分かる．また，$\hat{Ra}2$ 位以降のトピックについても見ると，トピックの順位は高くなっていることが分かる．

服飾，学校の局面に対して $\hat{Ra}1$ 位で結びつくトピックと，食事，趣味の局面に対して $\hat{Ra}1$ 位と $\hat{Ra}2$ 位で結びつくトピックについて，生起確率の高い上位 10 単語を抽出したものを表 5.3 に示す．服飾と学校については，表 1.1 に示したような局面の典型的な語や，その関連語が集まっていることが分かる．食事と趣味については，それぞれ 1 位と 2 位でやや傾向が異なるものの，こちらも表 1.1 に示したような単語が現れている．以上のことから，服飾，食事，趣味，学校の局面では，各局面を表現するために必要なトピックが，高い関連度で結びついていると考えられる．

これら 4 つの局面と比較して，地域の局面は高い F 値を示しているものの，$\hat{Ra}1$ 位から 4 位までに結びついてるトピックの順位が高くない．また，$\hat{Ra}1$ 位から 4 位までに結びついてるトピックである Topic62，Topic174，Topic152，Topic21 は，災害や行事，交通や気象といった局面とも高い関連度で結びついていることが分かる．これらのトピックについて分析するため，それぞれのトピックの生起確率の高い上位 10 単語を抽出したもの表 5.4 に示す．各トピック中の単語に注目すると，京都特有の単語が頻出していることが分かる．以上のことから，これらのトピックは京都という地域を表現するために必要なトピックであるため，地域の局面に対して高い関連度で結びついていると考えられる．

表 5.4 のいずれのトピックも，「京都」や「河原町」，「伏見」，「大阪」といった地名を意味する単語も出現していることが分かる．このことが，災害，行事，交通，気象の局面とも高い関連度で結びついている原因であると考えられる．例えば，災害の局面が付与される tweet とし

表 5.1　提案手法が局面を推定した tweet 例

正解局面	気象，健康
NBML が推定した局面	気象
TPE0 が推定した局面	気象，健康
tweet 本文	明日も関西は晴天で黄砂も飛んでこない様子． 最高気温も 28 ℃くらいまで上がるよう． 外出時には UV ケアもしっかりとして直射日光に気をつけなければ！ 熱中症にも気をつけないとね．水分補給を忘れずに…

表 5.2　提案手法が過剰に局面を推定した tweet 例

正解局面	学校，行事
NBML が推定した局面	学校，行事
TPE0 が推定した局面	学校，行事，地域
tweet 本文	【拡散希望】5 月 31 日 18 時から桂集会を開催します． 開催場所は桂川の近くです． 大学院試験のための勉強会やります．気軽に連絡ください．

て，地震の発生を報せるような tweet が例として挙げられる．地震の発生について言及する場合，震源地がどこであるかも記述されることが多い．行事についてはイベントの開催地，交通については駅名といった形で，これらの局面は地名をともなって言及されることが多い．以上のことが原因となって，表 5.4 に示すトピックと高い関連度で結びついたが，これらのトピック中では各局面を表現するために必要な単語の生起確率が低かったために，災害，行事，交通，気象の局面では F 値が低くなったと考えられる．

5.3　推定精度と対応関係に関する考察

図 4.2 より，学校の局面では $d \leq 11$ までは結びついているトピック数が 1 つであるが，適合率は上昇し，再現率は減少している．この原因として，他の局面に結びつくトピック数が増加しているために，tweet に対して算出されるスコアに変化が生まれ，推定する局面が変わっていることが考えられる．学校の場合では，他の局面のスコアが高くなり，学校と推定できる局面の数は減ったが，その分誤って推定していた tweet も少なくなったために，再現率が下がり，適合率が上がったと考えられる．

図 4.3 より，交通の局面ではトピック数の増加と共に，適合率は上がっている．また，$d = 6$ や $d = 11$ の結びつくトピック数が増加しているときに，再現率も大きく上がっている．$d = 6$

表 5.3　F 値が高い局面に対して結びつくトピックの上位語

トピック	上位語
Topic147 服飾 $\hat{Ra}$ 1 位	着る，浴衣，似合う，T シャツ，デート， スーツ，ピンク，衣装，シャツ，水着
Topic67 学校 $\hat{Ra}$ 1 位	勉強，テスト，終わる，試験，課題， 面接，集中，受験，受かる，合格
Topic132 食事 $\hat{Ra}$ 1 位	食べる，ご飯，美味しい，お昼，ごはん おいしい，アイス，寿司，ケーキ，チョコ
Topic154 食事 $\hat{Ra}$ 2 位	飲む，ビール，コーヒー，美味しい，お茶 呑む，飲める，ワイン，ジュース，おいしい
Topic4 趣味 $\hat{Ra}$ 1 位	見る，録画，テレビ，面白い，ドラマ， 様子，ビデオ，貞子，楽天，感動
Topic73 趣味 $\hat{Ra}$ 2 位	見る，感じ，見れる，途中，おもしろい 最後，最近，興奮，いい，コール

表 5.4　多くの局面に対して結びつくトピックの上位語

トピック	上位語
Topic62	京都，新聞，市内，イオン，宇治， 滋賀，会館，平成，タワー，住む
Topic174	京都，観光，河原町，交通，四条， 案内，鳥丸，便利，三条，地下鉄
Topic152	京都，美しい，神社，公園，咲く 商店，季節，散歩，伏見，綺麗
Topic21	京都，美山，体験，時代，着物 大阪，花魁，舞妓，大学，工房

表 5.5　交通の局面に対して結びつくトピックの上位語

トピック	上位語
Topic162 交通 $\hat{Ra}$ 4 位	乗る，バス，電車，新幹線，降りる， 着く，移動，帰り，向かう，夜行

のとき結びついているトピック数は 2 であり，$d = 11$ のとき結びついているトピック数は 4 であることが確認できる．表 4.6 より，交通の局面に $\hat{Ra}2$ 位と 4 位で結びついているトピックは，Topic174 と Topic162 である．また，Topic162 については $\sigma(\hat{Rt}(A,t))$ の順位が 12 位であることが確認できる．Topic174 において生起確率が高い単語ついて見ると，表 5.4 より主に京都の地名が頻出し，その中に「地下鉄」や「案内」といった交通に関連する単語が現れていることが分かる．このため，$d = 6$ で再現率が上がっていると考えられる．表 5.5 に，Topic162 において生起確率が高い上位 10 単語を示す．Topic162 については，上位の全ての単語が交通に関連する単語であることが分かる．このため，Topic162 の $\sigma(\hat{Rt}(A,t))$ の順位も高くなり，$d = 11$ で Topic162 と結びついたときに交通の再現率が大きく上がったと考えられる．

第 6 章

結論

本論文では，未知の tweet に対して適切な局面を推定するマルチラベリングを実現する，階層的推定法を提案した．提案手法は，教師なし学習と教師あり学習を組合せているところに特徴がある．第一段階では，教師なし学習で文書をクラスタリングできる LDA を用いて，大量の tweet からトピックを抽出する．第二段階では，少量のラベル付き tweet を用いて，トピックと局面の対応関係を構築する．未知の tweet に対して適切な局面を推定するため，トピックと局面の対応関係と，tweet から抽出した単語のトピック中の生起確率を用いて，局面毎にスコアを算出する．閾値を超えたスコアを持つ局面を，他の局面よりも相応しい局面として，未知の tweet に対して推定する．

提案手法の有効性を評価するため，京都市内で投稿された日本語 tweet を用いて評価実験を行った結果，未知の tweet に対して複数の適切な局面を推定できることを明らかにした．Naive Bayes をマルチラベル分類へ拡張した NBML と提案手法を比較した結果，NBML に比べて提案手法はより網羅的に局面を推定できることが明らかになった．トピックと局面の対応関係について分析した結果，F 値が高い局面では，その局面を表現するために必要な単語が集まったトピックが高い関連度で結びついていた．F 値が低い局面では，他の局面とも高い関連度で結びついているトピックが結びついていることを明らかにした．今後の課題は，トピックと局面の対応関係をより洗練することによって，F 値の向上を目指すことである．

謝辞

本研究を進めるにあたり，常に適切なアドバイスを行ってくださり，また共に助け合いながら頑張ってきた同期の山口裕太郎さん，大山鉄郎さん，堂前友貴さん，本当にありがとうございました．また，佐藤研究室の後輩である中岡義貴さん，川上未来さん，清野悠希さん，玉田雄基さんには研究室内外において非常にお世話になりました．共に佐藤研究室で学べたこと感謝しています．

関洋平助教授，若林啓助教授には，ゼミや研究セミナー等で研究のアプローチや技術的な相談などについてご指導頂き，修士研究を進める上で大変お世話になりました．池内淳准教授には，副指導教員として研究について熱心なご指導を頂きました．指導教員である佐藤哲司教授には，３年間もの間，熱心にご指導を頂き，研究のみならず多くのことを学ぶことが出来ました．心から感謝申し上げます．

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Machine Learning Research*, Vol. 3, pp. 993–1022, March 2003.

[2] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46, 1960.

[3] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning Research*, Vol. 29, No. 2-3, pp. 103–130, November 1997.

[4] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *The National Academy of Science*, Vol. 101, pp. 5228–5235, 2004.

[5] Yuki Hattori and Akiyo Nadamoto. Extracting tip information from social media. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services*, iiWAS '12, pp. 205–212, 2012.

[6] Kentaro Inui, Shuya Abe, Kazuo Hara, Hiraku Morita, Chitose Sao, Megumi Eguchi, Asuka Sumida, Koji Murakami, and Suguru Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In *Web Intelligence*, pp. 314–321, 2008.

[7] Wickre Karen. Celebrating #twitter7. https://blog.twitter.com/2013/celebrating-twitter7, Mar. 2013.

[8] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pp. 1137–1143, 1995.

[9] Taku Kudo. Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net/, 2005.

[10] Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka. Blog map of experiences: Extracting and geographically mapping visitor experiences from urban blogs. In *Proceedings of the 6th International Conference on Web Information Systems Engi-*

neering, WISE'05, pp. 496–503, 2005.

[11] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. Twiner: Named entity recognition in targeted twitter stream. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pp. 721–730, 2012.

[12] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] Endres Dominik Maria and Schindelin J E. A new metric for probability distributions. *IEEE Transactions on Information Theory*, Vol. 49, No. 7, pp. 1858–1860, 2003.

[14] Michael Mathioudakis and Nick Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pp. 1155–1158, 2010.

[15] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pp. 248–256, 2009.

[16] Martin Riedl and Chris Biemann. Topictiling: A text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, ACL '12, pp. 37–42, 2012.

[17] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 851–860, 2010.

[18] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, Wei Vivian Zhang, Deng Cai, and Xiaofei He. Whom to mention: Expand the diffusion of tweets by @ recommendation on micro-blogging systems. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pp. 1331–1340, 2013.

[19] Zhihua Wei, Hongyun Zhang, Zhifei Zhang, Wen Li, and Duoqian Miao. A naive bayesian multi-label classification algorithm with application to visualize text search results. *International Journal of Advanced Intelligence*, Vol. 3, No. 2, pp. 173–188, 2011.

[20] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pp. 261–270, 2010.

[21] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on*

Machine Learning, ICML '97, pp. 412–420, 1997.

[22] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. Auralist: Introducing serendipity into music recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pp. 13–22, 2012.

[23] Zhe Zhao and Qiaozhu Mei. Questions about questions: An empirical analysis of information needs on twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pp. 1545–1556, 2013.

[24] 池田佳代, 田邊勝義, 奥田英範, 奥雅博. Blog からの体験情報抽出. 情報処理学会論文誌, Vol. 49, No. 2, pp. 838–847, feb 2008.

[25] 山本雅人, 小笠原寛弥, 鈴木育男, 古川正志. 観光情報学:9. 東日本大震災時の twitter における情報伝播ネットワーク. 情報処理, Vol. 53, No. 11, pp. 1184–1191, oct 2012.

[26] 山本修平, 佐藤哲司. Twitter からの実生活情報の抽出法の提案. 第 4 回データ工学と情報マネジメントに関するフォーラム DEIM2013 論文集, F3-4, 2012.

発表論文

国際会議論文

- Shuhei Yamamoto and Testuji Satoh. Two Phase Extraction Method for Multi-label Classification of Real Life Tweets. Proceedings of the 15th International Conference on Information Integration and Web-based Applications & Services, iiWAS2013, pp. 16—25, 2013.
- Shuhei Yamamoto and Tetsuji Satoh. Two Phase Extraction Method for Extracting Real Life Tweets using LDA. Proceedings of the 15th Asia-Pacific Web Conference, APWeb2013, pp. 340—347, 2013.

国内査読付会議論文

- 山本 修平, 佐藤 哲司. 二段階抽出法を用いた実生活 Tweet のマルチラベル分類. 情報処理学会, マルチメディア，分散，協調とモバイルシンポジウム (DICOMO2013) 論文集, pp. 64—71, 2013.
- 山本 修平, 佐藤 哲司. 環境に適応する実生活情報の提示法, 情報処理学会, マルチメディア，分散，協調とモバイルシンポジウム (DICOMO2012) 論文集, pp. 266—273, 2012.

国内会議・研究会論文

- 山本 修平, 中岡 義貴，佐藤 哲司. 食材調理法の習得順に関する一検討. 電子情報通信学会, 信学技法, DE2013-38, pp. 31—36, 2013.
- 山本 修平, 佐藤 哲司. LDA を用いた実生活 Tweet の二段階抽出法. 第 5 回データ工学と情報マネジメントに関するフォーラム DEIM2013 論文集, C2-1, 2013.